Vienna 2019 Abstract Submission

Title
Wine Descriptions Provide Information: A Text Analysis

I want to submit an abstract for:
Conference Presentation

Corresponding Author
Bryan McCannon

E-Mail
bryan.c.mccannon@gmail.com

Affiliation
West Virginia University

Keywords
experience good; hedonic price; Latent Dirichlet Allocation; text analysis; topic modeling; wine

Research Question
Is there value in the text descriptions provided by expert critics that consumer's value?

Methods
Latent Dirichlet Allocation is used to quantify the topics of text descriptions of U.S. wines. A hedonic price equation is used with these topics used as explanatory variables.

Results
Descriptions of U.S. wine is evaluated. I show that once variety, region, and the numerical rating is properly controlled for, the text correlates with prices charged.

Abstract
Experience goods are a challenge for consumers. Frequently, they do not know how much utility they will receive if the purchase this good. This provides third parties the opportunity to provide a valuable service if they can provide some of that information to consumers. In wine, for example, professional expert critics provide numerical scores.

Additionally, though, it is common for these experts to provide a textual description. Presumably, this is informative as well. Economics research for the most part has ignored the added text. The only use of them is to create series of indicator variables capturing the existence of words the researcher thinks is important.

I use a computational linguistic algorithm, Latent Dirichlet Allocation, to measure the topics covered in textual descriptions of wine. Latent Dirichlet Allocation is an untrained, unsupervised topic modeling approach. The presumption is that when authors are writing on a similar topic, they tend to use the same words. Thus, the co-occurrence of words can be used to identify the topic covered in a document. With this method, each document can be assessed on the probability it falls within each potential topic. The method does not require the researcher's a priori knowledge of what are the important topics to be looking for. Thus, it is free from researcher bias.

With this methods in hand, I ask whether there is information in the text that consumers value. Wine is a prominent example of an experience good. There is substantial product differentiation in the market and consumers only have limited information on the utility they will receive when consumed. Thus, information is expected to be valuable.

It is not clear that they are. If the text is providing complementary information, such as food pairings, there is no reason to believe it is correlated with price. Also, it is cheap talk and can be thought of as babble. There is no cost or regulation to the words used. On the other hand, if the descriptions are providing information then some descriptions will be viewed favorable. That will lead to an increase in demand for that wine and, ultimately, increase its price. Other descriptions will receive consumers' unfavorable experiences and lead to that wine being of a lower price.

Evaluating descriptions of wine produced across the U.S., I use a hedonic price regression to explore whether the descriptions provide any new information not already available to the consumer. Hedonic price equations are a common way to evaluate consumer's preferences for product differentiated goods.

I first evaluate a sample of wine in New York state. Classifying the text descriptions into three topics, I show that the topics are correlated with price, until the varietal is controlled for. Looking at the words that have the greatest probability of falling within each topic, I show that the topic modeling is basically separating the white wines of the Finger Lakes region from the reds produced prominently in the Long Island region. Thus, the text is not providing any additional information that consumers value.

Next, I zone in one one particular varietal produced in one state: Oregon Pinot Noirs. Again, using LDA to evaluate the text descriptions provided by experts, I show that they are correlated with price. This persists until the numerical rating is also included as an explanatory variable. Thus, the text is simply telling consumers which wines are high quality and which are not.

As a final investigation, I consider one particular varietal, produced in one particular state, of one particular numerical rating. I choose 90 rated Cabernet Sauvignon wines of California. California is the dominant wine growing area of the U.S., Cabernet Sauvignon is a dominant varietal produced in the state, and the median and mode numerical rating provided by experts is 90. Considering this sample, I show again that the topics discussed in the description are correlated with price and do not lose their statistical significance when controls are included. Thus, indeed, there is informational value consumers value.

With this result in hand, I go back to the Oregon sample. Instead of only considering three topics, I consider ten topics. I show that once this is done, then three of the topics stand out as being correlated with price when all controls are included. Thus, a finer categorization of topics in that sample recovers the importance of some text descriptions. While not presented in the final paper, the same analysis can be done to the New York wine sample.

Thus, wine descriptions provide information that consumers value, which presumably shifts demand for wine, and results in prices.

The paper is the first to consider the relationship between the price of an experience good and text descriptions. While interesting to wine enthusiasts, the results suggest that other similar experience goods, such as movies, restaurants, cigars, etc. are likely influenced by the words used by people who comment and review them.

File Upload (PDF only)

- wine_aawe.pdf

Privacy

- Wine Descriptions Provide Information: A Text Analysis
- By using this form you agree with the storage and handling of your data by this website.

Wine Descriptions Provide Information
A Text Analysis

Bryan C. McCannon
West Virginia University*

08 February 2019

Abstract

I use a computational linguistic algorithm to measure the topics covered in textual descriptions of wine. I ask whether there is information in the text that consumers value. Wine is a prominent example of an experience good. There is substantial product differentiation in the market and consumers only have limited information on the utility they will receive when consumed. Thus, information is expected to be valuable. Evaluating descriptions of wine produced across the U.S., I use a hedonic price regression to explore whether the descriptions provide any new information not already available to the consumer. Initial results suggest that text descriptions are shown to lose their explanatory value when varietal and numerical ratings are included as controls. I then show that once the varietal, region, and numerical ratings are adequately controlled for, there is information in the descriptions that consumers value.

Keywords: experience good; hedonic price; Latent Dirichlet Allocation; text analysis; topic modeling; wine
JEL Codes: D83; L15; C81

1 Introduction

For a typical consumer wanting to buy wine, the number of choices is overwhelming. There are many varieties of wine, produced in countries around the world, across regions within a country, over time. Even within a particular varietal-region-vintage category, numerous producers market product differentiated wines.

To aid the consumer, wines are often reviewed by professionals. Typically, numerical scores are assigned ranging between zero and one hundred. These reviews are published in magazines and online web sites. Retailers also commonly provide the numerical ratings on the store shelves. An extensive literature has investigated whether the numerical ratings correlate with the price through, presumably, demand. The typical finding is that it does, but the correlation is not very strong. This suggests that there is more to a typical consumer's demand than the vertical product differentiation characteristics. Also, though, consumers are frequently given written descriptions in these reviews. These notes are provided by experts. Magazines and online sites publish these descriptions alongside the numerical score and other objective characteristics. Wine retailers not uncommonly provide the short text descriptions on the shelves next to the bottles.[1] I ask

*Department of Economics, PO Box 6025, Morgantown, WV 26506; bryan.c.mccannon@gmail.com. I appreciate comments from Orley Ashenfelter, Josh Hall, Brad Humphreys, Adam Nowak, and seminar participants at West Virginia University.

[1] In fact, it is quite common for producers to print descriptions directly on the bottle. An important distinction between these and the data analyzed here is that those descriptions are created by the seller, where the descriptions studied here are given by an expert who writes professionally on the wine market.

whether the text provides any additional information about the utility the consumer will receive from the wine above and beyond the quality score and the objective, observable characteristics (i.e., vintage, producer, variety, region).

Wine is just one example of an experience good where consumers' information is limited. Often in such markets third parties act to provide information. Text descriptions are a common way to do this. Thus, by evaluating wine descriptions, my work contributes to the broader question of the value of text in experience good markets.

There is reason to believe that the text is unrelated to a wine's price. For one, the descriptions may simply provide useful, complementary information such as food pairings or whether the wine needs cellaring. There is no reason to believe this necessarily correlates with price. The words used can reiterate information as well. To a typical consumer a description saying a wine has a *buttery* or another has a *cherry* flavor may be no different than simply telling her that the first one is a Chardonnay and the second one is a Pinot Noir. The text can be only giving typical characteristics of the wine's variety, which is already provided on the bottle.

Additionally, the descriptions can be viewed as cheap talk. Retailers and wine producers may simply select the descriptions that make the wine sound good. There is not necessarily a credible authority to the descriptions and the text is unregulated. Most any wine can be thought of as *wonderful*, *impressive*, or *interesting* (Especially if one is willing to stretch the truth!). In other words, there is not necessarily a cost to choosing which words to use in the description, so that Spence-type signaling cannot arise. If the text is treated as babble by consumers, then one would not expect a relationship between price and the description.

I explore the alternative hypothesis - - that there is useful information in a wine's textual description that consumers benefit from. The demand shift that is associated with an informative wine description is then reflected in its price. Descriptions that convey greater utility for the consumers lead to higher prices. Descriptions that clarify undesirable properties see price reductions. If the quality rating is a measure of vertical product differentiation (i.e., product quality), then the textual description can be thought of as providing information on the horizontal product differentiation. The empirical frustration is how to numerically analyze words in a rigorous, systematic way.

Early research attempting to do this basically created lists of indicator variables. The text is hand read and coded, sometimes by multiple independent research assistants. These indicator variables are then used in a hedonic price equation. For one example, Lecocq and Visser (2006) use simple indicator variables such as "excessiveness of acidity" (Yes or no?), "supple?" (Yes or no?), and "flat" (Yes or no?) to name just three of their variables. Advances in text analysis now allow researchers to systematically evaluate data sets of text. Using a free online wine review site, I use text analysis to categorize wine descriptions.

Specifically, I employ a topic modeling method known as Latent Dirichlet Allocation (hereafter LDA). Details of LDA are provided in Section 3. In short, though, LDA is an untrained, unsupervised algorithm. The set of documents is organized into T topics. Similar texts are grouped together to make up these

topics. Each document, then, can be scored on the probability it falls into each of these topics. Using this approach, I can classify wine descriptions into topics and ask if there is a particular type of description and an important set of words used in these descriptions that correlates with price.

Using descriptions of wines produced within the United States, I show that there is an important relationship. I first consider wines produced in New York state. I find that the description correlates with price charged, even when controlling for the numerical rating. This relationship disappears when observable traits of the wine are added as control variables. Exploring this effect, it turns out that the wine descriptions simply correlate with the wine's variety, basically distinguishing the Rieslings of the Finger Lakes region with the reds (e.g., Merlot) of the Long Island region.

Therefore, in a second data set I evaluate, within a particular variety, whether the text provides additional information explaining price. To do this, I consider data from reviews of Pinot Noir from Oregon. Here, again, the text is correlated with price and is robust to the inclusion of vintage, producer, and critic. The description's statistical significance diminishes, though, when the numerical rating is included as an explanatory variable. Thus, the numerical rating captures most of the information provided by the text.

At this point, it is not yet clear whether wine descriptions are useful. To further drill down in search of a potential relationship, I choose not only one varietal from one particular state, but also restrict to one particular quality rating. I consider a data set of 90-rated, California Cabernet Sauvignon. California is the dominant wine producing state in the United States, and Cabernet Sauvignon is its flagship varietal (or at least one of its leaders). Within this category, the median and mode numerical rating is 90. Considering only this subsample, I show that the text description again correlates with the wine's price. This time, though, it cannot be explained away by the other observable characteristics such as vintage, region, or critic. Thus, once the primary determinants of a wine's price is properly controlled for (i.e., region, varietal, and rating), the results are consistent with the hypothesis that the text descriptions do provide valuable information to the consumer affecting their demand and, ultimately, the price charged.

While the empirical investigation focuses on wine, the study of experience goods more broadly is problematic. We know little about how consumers acquire information in these settings. As stated, third-parties typically offer information. Other prominent examples include movies, restaurants, cigars, and music. In each, reviewers provide descriptions. Researchers have used the timing of reviews, such as when the Siskel and Ebert reviews are given (Reinstein and Snyder, 2005), to explore whether there is a demand effect. Unexpected opening weekend movie sales are used to identify social learning (Moretti, 2011). "Cold" releases of movies, unreviewed by critics, has been used to assess consumers' limited strategic thinking (Brown *et al.*, 2012). Of course, numerical ratings or other quantifications, such as the number of "stars", has a rich history of being studied in numerous experience markets (Jin and Leslie, 2003; Anderson and Magruder, 2012; McCannon, 2012). I provide the first rigorous analysis of textual descriptions and its influence on the prices of experience goods.

2 Literature Review

Numerous authors have explored the relationship between wine's price and its quality measured by numerical ratings provided by expert tasters (Oczkowski, 1994). For example, Goldstein *et al.* (2008) considers data from blind tastings and cannot find a relationship between price and the ratings given. For those with wine training, though, a positive relationship exists. Dubois and Nauges (2010) employ a structural approach to identifying experts' grades' relationship with price. In a field study experiment, Hilger *et al.* (2011) document a relationship between expert opinion and demand. Demand increases for high-scoring wine and falls for low-scoring ones. Consumers are poorly informed (Weil, 2007)[2] and demand experts' opinions (Ashenfelter and Jones, 2013). In fact, Ali *et al.* 2008) exploit an exogenous change in the timing of the release of a highly respected wine reviewer's numerical grading (Robert Parker) to document his influence on wine price. Thus, grading correlates with price.[3]

Regarding the text description, Combris *et al* (1997a; 1997b) and Lecocq and Visser (2006) collect not only jury grades of Bordeaux and Burgundy wines, but also have the experts tasting notes. As previously mentioned, indicator variables are created from these notes. They separate the sensory characteristics into olfactory (e.g., aromatic intensity, finesse of aromas, complexity) and gustatory features (firmness, excessive acidity, suppleness, flatness, fat, well concentrated, harmony, fine tannins, and finish). Each of these are either binary variables or ordered on a Likert scale. Combris *et al* (1997a) argue that it is the objective characteristics that correlate with price, while the sensory characteristics better correlate dwith thejury grade. Lecocq and Visser (2006) find evidence that some of the sensory variables have a statistically significant relationship with the wine's price, even controlling for the jury grade. Yang *et al.* (2009) uses a consumer panel of wine tastings. They have consumers complete a Likert scale evaluation ranging from 1 to 9 on eight dimensions, four related to the consumers preferences and four related to the observed intensity (aroma, flavor, astringency, bitterness). Trained and untrained tasters were considered. Ramirez (2010) counts the number of words used in tasting notes and shows that it correlates with price. Separating quality from producer reputation tends to suggest that reputation has a stronger relationship with price (Landon and Smith, 1998a; 1998b; Oczkowski, 2001).

Ashenfelter (2008) provides an important evaluation of prices of Bordeaux wine. He argues that most of the variation in prices can be explained by producer and vintage. The first-growth Bordeaux wines are famous and expensive. Those in the market for them can be expected to be quite informed. For less-famous wines, observable characteristics such as the producer and the vintage are available. Regional reputation matters for price (Schamel, 2009), as does producer technological investments (Gergaud and Ginsburgh, 2008). Weather in Napa is considered by Ramirez (2008). He finds that while weather correlates with

[2]Weil (2007) provides evidence that consumers cannot even distinguish wines. In his experiment, subjects were given three glasses of wine − two of which contained the same wine while the third was a different wine. Only 51% of the subjects who can tell the wines apart can match the tasters descriptions to the wines, which is not different from random.

[3]Relatedly, Crozet *et al.* (2012) evaluates the relationship between ratings of champagne producers and export data. They show that quality increases monotonically firm-level price.

quality ratings, it has a stronger correlation with price. My work complements their findings. I show that some of the unexplained variation in prices can be attributed to information consumers gather from the text describing the wine.

The use of LDA in economics is new and represents an important contribution to the field. The notable exception is the work of Hansen *et al.* (2018) who evaluates FOMC transcripts. They use the variation in topics discussed, assessed using LDA, as the dependent variable to appreciate how experience on the FOMC and transparency interact. This work differs in that I use LDA to gain valuable information to be used as an explanatory variable in the forecasting of a more-standard economic variable. Outside of economics, LDA is popular. As one measure, Blei *et al.*'s (2003) paper that introduced the algorithm has more than 25,000 citations. It has been used effectively in related fields such as political science. Grimmer (2010) uses LDA to analyze press releases from U.S. Senators. Quinn *et al.* (2010) use it to evaluate speeches in the U.S. Senate. It has been used recently in marketing to evaluate online discussions of products (Tirunillai and Tellis, 2014) and in accounting to identify trends in 10-K disclosures (Dyer *et al.*, 2017).

3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a computational linguistic algorithm (Blei *et al.*, 2003). It allows for the creation of communication measures based on topic models, which is a class of machine learning algorithms for natural language processing.

LDA allows for automatic clustering of any kind of text documents into a user chosen number of clusters, known as *topics*. It uses a probabilistic model of text data. The method's logic is that when authors write about a particular topic, they tend to use the same words. Hence, in texts about the same topic, similar words tend to co-occur. LDA describes each topic as a probability distribution over words, and each document as a probability distribution over topics.

I briefly describe the LDA algorithm. An interested reader is encouraged to consult Blei *et al.* (2003) or Schwarz (2018) for further details. Each document d in a set of documents D is described as a probabilistic mixture of T topics. A document topic vector, θ_d, describes the document. Rather, a document is determined by a probability distribution over topics. Each topic t in the set of topics T is described by a probability distribution over the vocabulary of V words present in all documents. Rather, within each topic, a probability distribution of words in the dictionary exists. For each document the topic proportions are drawn from a Dirichlet distribution with parameter α, and for every topic the word probability distribution is drawn from a Dirichlet distribution with parameter β. Thus, the researcher must only select the number of topics to organize the documents into and the two hyperparameters α and β.

Gibbs sampling is used to estimate the conditional probabilities that best explain the corpus of documents.[4] I follow the conventional norms in the literature by setting, as my baseline, $\alpha = 0.25$ and $\beta = 0.1$ (Schwarz, 2018). Furthermore, the baseline specifications in the paper consider organizing the documents

[4]Following Schwarz (2018), I use the `ldagibbs` command in Stata 15.

into three topics. I choose three because it is a coarse division and, therefore, is the topic differentiation that minimizes consumer's mental efforts. In a supplemental appendix, I provide alternative estimations considering more topics and show that the primary results are not sensitive to the use of three topics. Furthermore, I consider other values of the hyperparameters used and again show in the supplement that the mian results are not sensitive to the values chosen.

From this process, LDA calculates for each document the probability it falls within topic t, $\rho_d(t)$. With three topics, $t = 1, 2, 3$, it follows that $\rho_d(1) + \rho_d(2) + \rho_d(3) = 1$. These three variables, then, classify the topics discussed in the document. They become important explanatory variables in my analysis of wine prices.

LDA is not the only way to quantify text. Previous efforts have relied on *dictionary methods*. With these the researcher must first select a set of words that are believed to be important. For each document in the corpus, then, either the existence of the word or the number of words within the dictionary that arise are counted. See Gentzkow and Shapiro (2010) for a prominent example. This approach has a long history in wine reviews, as discussed previously, but has relied on hand coding of words researchers have thought is important. LDA, on the other hand, does not involve the researcher's pre-knowledge or discretion. It provides a way of uncovering hidden themes in text without having to link themes to particular word lists prior to estimation. King *et al.* (2017) points out that the human brain does not excel at recalling all keywords needed to adequately describe a topic. Instead, humans are good at making associations. Formal dictionary building techniques, such as LDA, do not rely on the researcher's ability to fully construct the keyword list. Thus, LDA is valuable when the researcher does not know a priori which words are the important ones to track. This allows me to avoid subjective judgments and to account for context.

As argued by Hansen *et al.* (2018) in their use of LDA, "we believe [our work] illustrates the value of combining traditional economic tools with those from the increasingly important world of Big Data for empirical research in economics". My work furthers this research agenda. To the best of my knowledge, mine is the first to use LDA to capture product differentiation and explain its relationship to price.

4 New York Wines

To investigate wine description's relationship with wine prices, I collect reviews from the web site winereviewsonly.com (accessed August 18, 2018). For each wine reviewed, the region of the state in which it is produced is provided. The varietal is listed. Price and the numerical rating are given. The identity of the reviewer, the date of the review, and the wine's vintage are also published. Importantly, a textual description is given.

4.1 Data

Within this data set, I first consider wine produced in New York state. There are 124 wines. New York has two dominant wine regions. The Finger Lakes region of upstate New York specializes in the growing

Table 1: New York Wine

	Price (avg.)	Rating (avg.)	# of Wines	Finger Lakes	Long Island	Other
White Wine						
Riesling	18.85	90.84	59	56	3	0
Gewürztraminer	18.57	89.71	7	7	0	0
Grüner Veltliner	15.00	93.20	5	5	0	0
Vidal Blanc	24.25	91.25	4	4	0	0
Chardonnay	18.75	89.75	4	1	3	0
Rkatsiteli	23.33	90	3	3	0	0
Blanc de Blanc	25.00	90.67	3	3	0	0
Sauvignon Blanc	22.00	89.67	3	1	2	0
Chenin Blanc	26.00	90	3	0	3	0
white blend	21.22	90.67	9	3	4	2
Red Wine						
Merlot	39.29	89.71	7	0	7	0
Cabernet Franc	30.50	89.83	6	2	4	0
red blend	23.38	91.00	8	3	1	4
Rose	19.00	89.67	3	1	2	0

The blend categories also include other varietals with two or fewer observations.

of Riesling and other central European white wines. Also, Long Island has a niche wine industry as well. Production is divided between reds (48%) and whites (52%). Overall, 72% of the wines reviewed come from the Finger Lakes and 23% come from Long Island. The following table provides a list of the wine varieties in the sample.

Thus, there is variation within a region and variation across regions in the wines produced. White wines tend to have higher numerical ratings, led by Grüner Veltliner, but red wines tend to have higher prices, led by Merlot. Riesling is the most popular wine in the data set making up almost 48% of the observations. This is the wine of choice in the Finger Lakes region. For the full sample, the average price is \$21.50 ($\sigma = 9.05$ with a minimum, median, and maximum of 8, 18, and 60, respectively). The numerical ratings average 90.63 (with $\sigma = 2.14$ and a minimum, median, and maximum of 86, 90, and 96, respectively).

First, for each description I remove basic punctuation.[5] Without doing so, LDA would differentiate "wine", "wine,", and "wine." as separate words, for example. I do not remove any other words from the descriptions. Since one of the values of this method is that the analysis is untainted by researcher influence, I choose to manipulate the text as little as possible.

As mentioned, my baseline evaluation is to organize the descriptions into three topics. My motivation is that considering only three topics allows for the simplest and coarsest differentiation for consumers. If a particular topic can be shown to be highly correlated with price when documents are organized into only

[5]Specifically, the period, question mark, and exclamation marks at the end of each sentence are removed. Also, the single and double quotation marks, the double hyphen, dash, and parentheses are eliminated. I do leave the hyphen (to allow for hyphenated words to be treated as single units), dollar sign, and the percentage symbol. I do not spell check the text or engage in any manipulation. My light touch approach is chosen to emphasize the unsupervised value to the topic.

Table 2: Explaining Prices of New York Wines

	(1)	(2)	(3)	(4)	(5)
Rating		0.028 *	0.056 ***		0.080 ***
		(0.016)	(0.015)		(0.021)
Topic 1	0.092		0.161	0.117	0.048
	(0.127)		(0.122)	(0.218)	(0.200)
Topic 2	0.557 ***		0.706 ***	-0.030	0.013
	(0.122)		(0.122)	(0.290)	(0.266)
Vintage Fixed Effects?	No	No	No	Yes	Yes
Varietal Fixed Effects?	No	No	No	Yes	Yes
Region Fixed Effects?	No	No	No	Yes	Yes
Critic Fixed Effects?	No	No	No	Yes	Yes
$F(\text{T1} = \text{T2} = 0)$	11.13 ***	–	17.48 ***	0.20	0.03
R^2	0.159	0.027	0.251	0.640	0.703
AIC	87.1	102.7	75.1	69.9	49.7
N	121	121	121	117	117

Dependent variable = $log(price)$.
Standard errors presented in parentheses; *** 1%, ** 5%, * 10% level of significance.
There are 16 vintage, 16 varietal, 4 region, and 12 critic fixed effects included.
A constant is included in each specification, but not reported.

three topics, then allowing for a finer classification should remove any subtopics unrelated to price and enhance the estimated effect. Later, I check the result's sensitivity to the number of topics selected. Thus, for each observation, three new variables are created: *Topic 1*, *Topic 2*, and *Topic 3*. Each ranges between 0 and 1, they sum to one, and can be interpreted as the probability the description falls in each topic.

4.2 Hedonic Price Estimation

To evaluate the relationship between the description and price, I estimate a hedonic price equation. The dependent variable is the log of the price. Specifications differ in which control variables are included. The following table provides the results.

The first specification, (1), suggests that Topic 2 is associated with higher prices, relative to Topic 3, the omitted one. As to be expected, the numerical rating is associated with a higher price, (2). Combining the specifications, the second topic for the text descriptions maintains its statistical significance, (3).

These do not control for the wine's characteristics. Columns (4) and (5), then, include other observable characteristics of the wine; namely the vintage, varietal, region, and critic evaluating the wine. Including these controls, the wine's description loses its statistical significance. The difference between the wine descriptions disappears, as well as the collective explanatory value, as indicated in the F-stat for the joint hypothesis that the topics all have a zero effect, $F(\text{T1} = \text{T2} = 0)$. Including these controls dramatically

Table 3: Varietal and Descriptions

	(1)	(2)	(3)	(4)
Topic 1	0.048	0.113	0.117	0.099
	(0.129)	(0.130)	(0.196)	(0.125)
Topic 2	0.550 ***	0.255	0.523 ***	0.296 **
	(0.141)	(0.175)	(0.166)	(0.147)
Vintage Fixed Effects?	Yes	No	No	No
Varietal Fixed Effects?	No	Yes	No	No
Region Fixed Effects?	No	No	Yes	No
Critic Fixed Effects?	No	No	No	Yes
$F(\text{T1} = \text{T2} = 0)$	8.4 ***	1.2	2.1	5.9 ***
R^2	0.306	0.337	0.295	0.228
AIC	88.8	88.3	82.7	87.7
N	117	121	121	121

Dependent variable = $log(price)$.
Standard errors presented in parentheses; *** 1%, ** 5%, * 10% level of significance.
There are 16 vintage, 16 varietal, 4 region, and 12 critic fixed effects included.
A constant is included in each specification, but not reported.

increases the goodness of fit as well. This suggests that the wine's description is not providing any more information than the basic information of who produced the wine and what type of wine it is. The numerical rating, on the other hand, continues to provide additional information explaining the price charged.

Since the fixed effects as a group take away the explanatory power of the text description, the following table adds one set of fixed effects at a time to explore which are causing this to occur.

Topic 2 remains positive and statistically significant in all specifications except (2). Furthermore, the topics are jointly zero there as well. In (2), only the varietal is controlled for. This suggests that the text description is providing the consumer with the same information as the wine's variety. Once it is controlled for, there is no additional explanatory effect on price.

Up to this point, only the generic labels have been used to identify the topics covered in the text. Table 4 provides a breakdown of some of the popular words that comprise each of the three topics. To illustrate, ten of the top twenty words are provided. Within each of the three columns, the first number is the probability weight put on the word. That is, I present the probability of observing that particular word (from the set of all words in the vocabulary) conditional on that topic being selected. The number in the parentheses is the word's ranking.

Topic 1, which I label as *White Fruit*, consists primarily of fruit references that are typically associated with Rieslings and other similar German/Austrian white wines. The second category, labeled *Red Fruits*, explicitly names Cabernet Franc and Merlot, and uses darker words such as "black", "cherry", and "acidity". The final topic, labeled *Awards*, consists of words that are associated with wine competitions. Table 1 shows

Table 4: Text Analysis: New York Wine

Topic 1 **"White Fruit"**		Topic 2 **"Red Fruit"**		Topic 3 **"Awards"**	
riesling	0.0281 (1)	fruit	0.0175 (2)	international	0.0188 (2)
apple	0.0196 (5)	shows	0.0120 (4)	competition	0.0164 (3)
fruit	0.0194 (6)	black	0.0119 (5)	platinum	0.0158 (5)
peach	0.0160 (7)	cabernet	0.0083 (9)	challenge	0.0140 (7)
rieslings	0.0131 (8)	merlot	0.0077 (10)	award	0.0121 (8)
lemon	0.0127 (9)	cherry	0.0077 (12)	flavor	0.0107 (10)
fruits	0.0111 (11)	acidity	0.0071 (13)	minerality	0.0104 (12)
white	0.0092 (14)	franc	0.0070 (14)	aromas	0.0099 (13)
honey	0.0089 (16)	balanced	0.0069 (15)	flavors	0.0092 (15)
citrus	0.0083 (20)	sauvignon	0.0056 (17)	bright	0.0079 (18)

that the German/Austrian whites (Riesling, Gewürztraminer, and Grüner Veltliner) have above average numerical ratings and below average prices. The reds (Merlot and Cabernet Franc) have below average ratings but above average prices. Therefore, it is not surprising that *Red Fruit* is associated with a higher price in the regression analysis, even controlling for the numerical rating, until the varietal is controlled for.

It is important to note that a word can appear in multiple lists. The numbers provided in the table are the probability that the particular word arises in a document, given that the document is one on that topic. In fact, every word in the vocabulary would have such an estimated probability for each topic. Table 4 simply identifies a subset of words with the greatest weights. It may be more informative to consider the probability that a document is on a particular topic given the existence of that particular word in the document. For example, the probability the word "peach" is used in a document that is within the White Fruit topic is 1.6% (Table 4). On the other hand, using its weight in the other two topics, the probability a document that uses the word peach is one on the White Fruit topic is 84.7%.

A few samples should shed light on how LDA classifies texts here. Using the estimated values of $\rho_d(t)$, those descriptions in the top deciles are those that can be categorized clearly within a particular topic. For example, an observation of a description that falls within the Light Fruit topic is provided below.

> The **vineyards** of New Yorks ***Finger Lakes*** are **proving** to be a **great source** of high **quality** ***Rieslings*** The 2014 **Hermann** J **Wiemer** Dry ***Riesling*** **makes** a **striking impression** with its ***fruit*** **intensity** and ***shows*** the **beautiful complexity** and racy **appeal** of the **worlds** best ***Rieslings*** The **aromas** are full and **forward** with ***lovely*** **scents** of ***peach*** **apricot grapefruit** and **Meyer** ***lemon fruits*** **enhanced** by **nuances** of ***white flowers*** and ***honey*** The ***flavors*** are pure and **complex** with the ***citrus*** and ***peach fruits*** **followed** by ***honey*** and ***spice*** **tones** With its ripe **style intensity** of ***fruit*** and **electric** ***acidity*** it is **reminiscent** of a trocken ***Riesling*** from **Germanys Pfalz** ***region*** It is an **outstanding example** of ***Finger Lakes*** **quality** and well **worth seeking** out

The bolded words are those that arise in the top 200 words for the topic, and the italicized words are those that show up in the top 25. As stated, punctuation has been removed from the text. As one can see,

numerous words just from the ten posted previously show up in the text. This particular wine is priced at $24 and received a 91 rating, which puts it right at the mean price and rating for wines in the sample. This is consistent with the findings in Table 3 that the Light Fruit category is not statistically different.

The Red Fruit topic is shown to have a price premium (Table 2), until the varietal is controlled for (Table 3). As with the White Fruit, samples of descriptions within the top decile of those texts that have the highest value of *Topic 2* are presented.

> Great Cab ***Franc*** from the Finger Lakes who knew Not me before tasting this wine but I found it entirely **convincing** and **actually** well worth the hefty price tag Deep and dark it displays **blackberry** ***fruit*** with a ***black cherry*** note as well and the wood and **tannin** are **nicely** weighted in **balance** with the ***fruit*** A ***light*** **herbal** aroma is true to the **grape** ***variety*** and is **quite pleasant** ***without*** seeming **green** or under-ripe

This Cabernet Franc is explicitly named as one in the text and uses many descriptive words that show up as important in defining the Red Fruit topic.[6] Interestingly, while this one is rated a 90, it is priced at $40, which is more than two standard deviations above the mean price in the sample. As another example,

> Deeply **flavorful** with good intensity that is **achieved** ***without*** a lot of body or **weight Red** and ***black*** **berry** ***notes*** are very **pleasant** and **tannin** and wood are **nicely** ***balanced***

This Merlot from the North Fork region of Long Island has a short description, but is packed with descriptive words indicating it is a red wine. This wine received a numerical rating of only 88, but has a price of $60, which is actually the most expensive wine in the sample. Without controlling for observable characteristics, Red Fruit wines have above average prices. Once the varietal is controlled for, the descriptions simply provides information on this variety so that there is no residual value to the description left.

The Awards topic, on the other hand, is shown to come with statistically lower prices (Table 2). Again, looking at a description from the top decile of those with the highest probability of being within the Awards topic illustrates the description's fit.

> The Blue Waters ***Riesling*** is yet **another superb** expression of the **grape** from **Swedish Hills Winery** The 2013 **offers richness** and structure with ***notes*** of mandarin orange and **tropical** ***fruit Platinum award winner*** at the 2015 ***Critics Challenge International*** Wine ***Competition***

This description consists almost exclusively of words that arise in the Awards topic. In fact, the words ranking 2, 3, 4, 5, 7, 8, and 9 in the Awards topic list are all included.[7] Overall, for this observation, the probability it falls within the Awards topics is 83.4%. Interestingly, it received a 91 rating as well, but only costs $14. Thus, it is offered at a price discount.

As another example,

[6] In fact, thirteen additional words from this description show up in the top 1000 words describing the topic.

[7] Additionally, "structure" (#235), "mandarin" (#212), "orange" (#288), and "expression" (#383) are also words arising commonly in this topic and show up in this particular text.

> **Swedish Hills** Humphreys Vineyard ***Riesling*** **offers** ***notes*** of wet ***stone*** and **tropical** ***fruit*** with good **minerality** all in a ***beautifully*** **balanced package** that is made for raw or **steamed** shellfish **Another example** of the ability of the ***Riesling*** **grape** to thrive in **upstate** New York ***Platinum award winner*** at the 2015 ***Critics Challenge International*** Wine ***Competition***

Again, a text description full of words used in the Awards topic arise. This entry has a value for *Awards* of 0.8374. It was given a numerical rating of 95, putting it in the upper echelon of wines in the sample (more than two standard deviations above the mean), but is priced at only $22. Thus, text descriptions that fall within the Awards topic are priced less. Recognizing, though, that both of these examples are Rieslings, which are on average lower in price (Table 1), controlling for varietal and the numerical rating eliminates the informativeness of these reviews.

5 Oregon Pinot Noir

The results from the analysis of New York wines show that the wine descriptions are not providing much more information that what knowledge of the wine's variety provides. To further explore the text's informational value, I isolate one popular varietal and evaluate differences in the wine's description and whether a systematic relationship with the price exists. Furthermore, because a wine variety tends to differ with the region in which it grows, I focus on one specific varietal in one specific state. I choose the Pinot Noirs grown in Oregon.

There are 238 Pinot Noir wines produced in Oregon within the sample. As with the New York wines, I first conduct an LDA analysis on the wine descriptions maintaining the baseline of three topics. As before, punctuation is dropped from the text. Also, here, the words "pinot", "noir", and "Oregon" are removed (along with plural and possessive forms). They are removed because these three words are true for every wine in the sample and, therefore, is redundant for a critic to include in his/her review. Categorization of the descriptions based on the use of one of these words would be uninformative to the consumer. Table 5 provides a sample of the words associated with each of the three topics, along with the estimated probability the word arises for a description that falls within the topic and the ranking of the word's importance in the topic's classification (provided in the parentheses).

I provide a descriptive label for each of the three categories. As one can see, the first topic, *Light Fruit* consists of words such as "light", "subtle", "fresh", and "delicate". These descriptions relate to the soft tastes and smells of the wine. The second topic, *Estate Reserve*, are descriptions to the terroir and winemaker's status. The third topic, *Flavorful Fruit*, consists primarily of references to berries, flowers, and herbs that describe the wine's taste and smell. These simple descriptive names will be used for the three topics.

Notice that the word fruit shows up as important in two topics. This illustrates the importance of LDA. If keywords are select beforehand by the researcher, it would be reasonable for him/her to classify fruit flavors together. Here, LDA reveals that such a division is not optimal. Berry related fruit should be separated from light, delicate fruit flavors and smells. Thus, it is the co-occurrence of words that matters, not the existence

Table 5: Text Analysis: Oregon Pinot Noir

Topic 1 **"Light Fruit"**		Topic 2 **"Estate Reserve"**		Topic 3 **"Flavorful Fruit"**	
fruit	0.0264 (1)	vineyard	0.0142 (3)	cherry	0.0317 (1)
notes	0.0146 (2)	bottling	0.0107 (5)	fruit	0.0210 (3)
light	0.0104 (3)	estate	0.0102 (6)	flavors	0.0170 (4)
subtle	0.0084 (6)	valley	0.0095 (7)	finish	0.0168 (5)
fresh	0.0079 (9)	reserve	0.0072 (11)	palate	0.0161 (6)
little	0.0078 (10)	williamette	0.0069 (12)	fruits	0.0140 (8)
flavor	0.0073 (13)	vineyards	0.0069 (13)	raspberry	0.0106 (12)
sweet	0.0064 (15)	hills	0.00671 (15)	vanilla	0.0097 (13)
delicacy	0.0062 (16)	complexity	0.0059 (16)	blackberry	0.0094 (14)
delicate	0.0055 (19)	earthy	0.0054 (19)	texture	0.0088 (16)
balance	0.0048 (26)	winemaker	0.0048 (23)	strawberry	0.0069 (22)
sweetness	0.0042 (31)	domaine	0.0043 (28)	floral	0.0048 (34)
structure	0.0039 (36)	elegance	0.0036 (36)	berry	0.0048 (35)
flavors	0.0037 (39)	character	0.0036 (37)	mushroom	0.0038 (44)
bright	0.0032 (45)	elegant	0.0031 (49)	herbs	0.0037 (45)

Table 6: Descriptive Statistics: Pinot Noir

	μ	σ	min.	med.	max	N
Price	42.02	18.82	15	38	150	237
Rating	90.85	2.30	85	91	96	236
Light Fruit	0.3549	0.2964	0.0041	0.2559	0.9677	238
Estate Reserve	0.2375	0.2667	0.0043	0.1226	0.9656	238
Flavorful Fruit	0.4076	0.3062	0.0036	0.4024	0.9856	238

of a single word. Observing the word fruit in a document means that the probability that document is on the first topic is only 53.7% even though it is the word with the greatest probability of being used, given that the document is on the first topic. Table 6 provides descriptive information about their price and quality rating for this data set.

Oregon Pinot Noirs have substantially higher prices than New York wine with more variation in price. Ironically, the numerical rating is not much greater. The Flavorful Fruit topic is most prevalent in the text descriptions in that the mean value placed on the probability of being within this topic across all documents is greatest.

Investigating one specific varietal in one specific region of the world, does the wine's description have a relationship with its price? The following table provides results from a hedonic price equation with log price as the dependent variable.

As before, the numerical rating is highly correlated with price, (1). An increase in rating by one standard deviation increases the price at the mean by approximately 22%.

Table 7: Explaining Prices of Oregon's Pinot Noir Wine

	(1)	(2)	(3)
Rating	0.095 ***		0.090 ***
	(0.010)		(0.011)
Light Fruit		-0.381 **	-0.297 *
		(0.184)	(0.156)
Estate Reserve		0.441 ***	0.118
		(0.194)	(0.170)
Vintage Fixed Effects?	Yes	Yes	Yes
Producer Fixed Effects?	Yes	Yes	Yes
Critic Fixed Effects?	Yes	Yes	Yes
$F(\text{LF} = \text{ER} = 0)$	–	3.55 **	2.80 *
R^2	0.652	0.527	0.662
AIC	247.4	181.0	103.9
N	235	237	235

Dependent variable = $log(price)$.
Standard errors presented in parentheses; *** 1%, ** 5%, * 10% level of significance.
There are 16 vintage, 22 producer, and 14 critic fixed effects included.
A constant is included in each specification, but not reported.

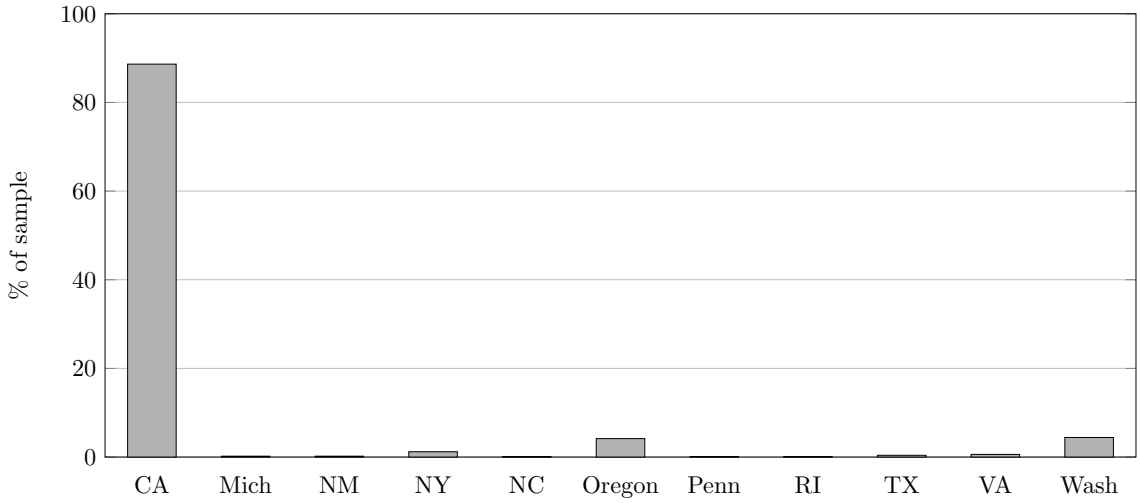
The text descriptions also exhibit a statistically significant relationship with the price, (2). Even controlling for vintage, producer, and the wine critic, (3), the *Light Fruit* category sees lower prices and the *Estate Reserve* has higher prices than the *Flavorful Fruit* omitted category. Additionally, the F-statistic for the joint hypothesis that the topics are all zero is large, $F(\text{LF} = \text{ER} = 0) = 3.55$. This suggests that the text description provides information to consumers.

The final column calls this finding into question. Here, both the text description and the quality rating are included (along with all controls). The description variables's magnitude reduces substantially. Also, the statistical significance is dampened. Thus, the numerical rating captures the important differences that drive wine prices. This finding can be explained by the text descriptions correlating with the numerical rating. The rating and vintage, producer, and critic controls explain 65% of the variation in price. Adding the text description increases this to only 66%.

6 California Cabernet Sauvignon

To further drill down into the relationship between a wine's textual description and its price, I now isolate not only one particular varietal in one particular state, but also zone in on one particular numerical rating. In the United States, California dominates the wine industry. The data set obtained, for example, contains far more California wines than any other state. Figure 1 illustrates.

Figure 1: Wines Reviewed by State



Each column depicts the proportion of the sample represented by that state.

Within California, Cabernet Sauvignon is popular. It makes up 17.8% of the reviews.[8] Within California Cabs, the mean, median, and mode numerical rating is a 90. This occurs for 259 wines. Therefore, I isolate 90-rated, California Cabernet Sauvignon reviewed.

First, the descriptive statistics of the subsample are described in Table 8.

Of this specific cohort, more than one-half of the wines are from the Napa region, which is to be expected given its prominence in the production of Cabernet Sauvignon. Regarding price, the substantial variation is noteworthy. Each of these 259 wines are rated by wine critics at exactly the same numerical score. Prices, though, range from only $8 (Oak Grove Family Reserve) to $150 (Chateau St. Jean).

As before, I organize the documents into three topics. Table 9 explores the relationship between these topics and the wine's price.

Again, the topics created via LDA are highly correlated with the wine's price. This time, though, the statistical significance remains when all control variables are included. For the sample of 90-rated, California Cabernet Sauvignons, the vintage, region of the state, and critic fixed effects are included.

Thus, wine descriptions do provide additional information that consumers value explaining over 12% of the variation in the wine prices.

Topics 1 and 2 experience a price premium relative to Topic 3, the omitted one. In Table 10, I provide information on the words used that make up each topic.

The top panel provides the top twenty words. The bottom panel highlights five additional words that help label each category. I label the first topic *Structure* as it consists of words related to specific dimensions that

[8]This is exceeded only by Pinot Noir (20.3%), and the only other varietals nearing Cabernet Sauvignon's popularity are Chardonnay (14.4%), Sauvignon Blanc (7.0%), Merlot (5.0%), and Zinfandel (4.9%). Of course, red blends & meritage (8.7%) and white blends (1.8%) are popular as well.

Table 8: Descriptive Statistics: 90-rated, California Cabernet Sauvignon

	μ	σ	min.	med.	max
Price	44.19	24.32	8	40	150
Topic 1	0.2802	0.2997	0.0081	0.1528	0.9853
Topic 2	0.2544	0.2479	0.0043	0.1497	0.9626
Topic 3	0.4654	0.3106	0.0061	0.4924	0.9784
Napa	0.5096	0.5009	0	1	1
Sonoma	0.1877	0.3912	0	0	1
Alexander Valley	0.0958	0.2949	0	0	1
Paso Robles	0.0881	0.2840	0	0	1
Dry Creek Valley	0.0345	0.1828	0	0	1

The omitted region is all other areas of California.

Table 9: Explaining Prices of 90-Rated, California Cabernet Sauvignon Wine

	(1)	(2)	(3)	4)	(5)
Topic 1	0.700 ***	0.405 ***	0.503 ***	0.554 ***	0.393 **
	(0.115)	(0.103)	(0.105)	(0.204)	(0.181)
Topic 2	0.259 *	0.206 *	0.154	0.582 ***	0.335 **
	(0.139)	(0.120)	(0.125)	(0.178)	(0.161)
Vintage Fixed Effects?	No	Yes	No	No	Yes
Region Fixed Effects?	No	No	Yes	No	Yes
Critic Fixed Effects?	No	No	No	Yes	Yes
$F(\text{T1} = \text{T2} = 0)$	18.49 ***	7.84 ***	11.67 ***	6.46 ***	3.11 **
R^2	0.126	0.460	0.334	0.299	0.556
AIC	396.7	306.8	336.4	361.7	288.2
N	259	258	259	259	258

Dependent variable = $log(price)$.
Standard errors presented in parentheses; *** 1%, ** 5%, * 10% level of significance.
There are 19 vintage, 6 region, and 12 critic fixed effects included.
A constant is included in each specification, but not reported.

Table 10: Text Analysis: 90-rated, California Cabernet Sauvignon

	Topic 1 **"Structure"**		Topic 2 **"Style"**		Topic 3 **"Dark Fruit"**	
1.	fruit	0.0194	valley	0.0199	fruit	0.0310
2.	flavors	0.0183	tannins	0.0174	black	0.0168
3.	tannins	0.0179	vintage	0.0135	spice	0.0153
4.	alcohol	0.0175	fruit	0.0101	finish	0.0134
5.	blackberry	0.0129	supple	0.0090	flavors	0.0125
6.	color	0.0124	which	0.0089	price	0.0106
7.	valley	0.0122	their	0.0082	aromas	0.0099
8.	french	0.0117	always	0.0071	years	0.0096
9.	blend	0.0115	vineyard	0.0067	notes	0.0086
10.	black	0.0114	flavors	0.0061	cherry	0.0085
11.	notes	0.0110	alexander	0.0055	cassis	0.0082
12.	finish	0.0107	makes	0.0055	tannins	0.0079
13.	months	0.0098	rather	0.0053	blackberry	0.0076
14.	spice	0.0097	maker	0.0053	sweet	0.0070
15.	length	0.0084	power	0.0052	would	0.0068
16.	vineyards	0.0080	theres	0.0051	flavor	0.0068
17.	bright	0.0076	complex	0.0050	vanilla	0.0067
18.	acidity	0.0071	because	0.0050	balance	0.0060
19.	balanced	0.0069	drink	0.0049	balanced	0.0059
20.	aromas	0.0068	example	0.0049	offers	0.0058
Others						
	layered	0.0067 (23)	complex	0.0050 (17)	cherries	0.0054 (23)
	texture	0.0066 (24)	polished	0.0037 (29)	chocolate	0.0048 (29)
	refined	0.0062 (29)	juicy	0.0036 (32)	depth	0.0047 (30)
	shows	0.0051 (33)	style	0.0032 (39)	character	0.0045 (34)
	structure	0.0049 (35)	lively	0.0032 (43)	dried	0.0043 (37)

the critic is writing about: Flavors, alcohol, color, notes, finish, length, etc. The sensory characteristics of the wine, then, coincide with a price premium paid by consumers. I label the second topic *Style*. Descriptive words such as "supple", "power", "complex", and "polished" are summary words used to convey the wine's properties. The final topic, which I label as *Dark Fruit*, consists of words like "spice", "cherry", "blackberry", and "chocolate". Wines that have descriptions that fall into the Dark Fruit topic tend to be sold at a price discount, even conditioning on the quality rating, varietal, state, region within the state, and vintage.

In a supplemental appendix, I redo the analysis on California Cabernet Sauvignon wines, but consider other popular numerical ratings. The results extend to these complementary analyses. Some topics pop as being sold at a price premium, relative to another topic that is sold at a relative price discount. Across these other numerical ratings the significance of the text descriptions remains when vintage, region, and critic fixed effects are included.

7 Oregon Reconsidered

The observation in Section 5 is that, using Oregon's Pinot Noirs, the textual descriptions lose their statistical significance when the numerical rating is included. The punchline from Section 6, using California's Cabernet Sauvignons, is that indeed there is a relationship once the numerical score is properly controlled for. Presumably, these two findings complement one another because only three topics were allowed. This forces potentially highly differentiated products into the same category. If more topics are created, then the separation should allow for the text descriptions to maintain their importance.

To explore this, I return to the Oregon subsample. Rather than bucket the wine descriptions into three topics, I instead allow there to be ten different topics covered in the descriptions.

With these finer categorizations, I explore whether the price of Pinot Noir wines made in Oregon is related to the wine's description. If it is the case that some descriptions are informative, but organizing the wines into three topics does not allow for the valuable information to be differentiated, then some topics should remain correlated with price when the numerical rating is included. Table 11 extends Table 7 previously presented.

Compared to Topic 10, the omitted one, Topic 1 and Topic 2 maintain a statistical difference with the former associated with lower prices and the latter coming with a price premium. Topic 1 uses words related to the freshness of the wine (e.g., "style", "fresh", "herbs", and "purity"), Topic 2 continues with the Estate Reserve words such as "reserve", "delivers", "domaine", "savory", and "quality". Thus, even within this sample, there is a relationship between the text descriptions and the price.

Also, while not presented here, the New York sample can be reconsidered allowing for more topics. As is the case with Oregon wines, some topics become statistically significant once the documents are more finely subdivided. Thus, evidence exists in all samples that price correlates with the textual descriptions.

Table 11: 10-Dimension Text Descriptions of Oregon Pinot Noir

	(1)	(2)	(3)	(4)
Rating		0.101 **		0.090 ***
		(0.010)		(0.010)
Topic 1	-0.848 **	-1.089 ***	-0.754 *	-0.650 *
	(0.398)	(0.353)	(0.407)	(0.355)
Topic 2	0.002	-0.328	1.032 ***	0.664 *
	(0.306)	(0.286)	(0.396)	(0.354)
Topic 3	-0.227	-0.429	0.727 *	0.559
	(0.282)	(0.262)	(0.394)	(0.343)
Topic 4	-0.335	-0.184	-0.3‘0	-0.221
	(0.296)	(0.274)	(0.263)	(0.241)
Topic 5	-0.749 *	-0.731 ***	0.074	-0.219
	(0.397)	(0.356)	(0.38)	(0.341)
Topic 6	-0.414	-0.617 **	-0.367	-0.333
	(0.285)	(0.266)	(0.321)	(0.286)
Topic 7	0.154	-0.131	0.732 **	0.517
	(0.330)	(0.302)	(0.355)	(0.314)
Topic 8	-0.288	-0.325	-0.295	0.015
	(0.321)	(0.294)	(0.360)	(0.320)
Topic 9	-1.028 ***	-1.137 ***	-0.459	-0.513
	(0.368)	(0.330)	(0.352)	(0.312)
Vintage Fixed Effects?	No	No	Yes	Yes
Producer Fixed Effects?	No	No	Yes	Yes
Critic Fixed Effects?	No	No	Yes	Yes
R^2	0.136	0.394	0.575	0.707
AIC	239.8	157.2	169.8	84.5
N	237	235	237	235

Dependent variable = $log(price)$.
Standard errors presented in parentheses; *** 1%, ** 5%, * 10% level of significance.
There are 16 vintage, 22 producer, and 14 critic fixed effects included.
A constant is included in each specification, but not reported.

8 Conclusion

In markets for experience goods, consumers presumably are in need of information. Text descriptions can provide this. Using wine as a prominent example of an experience good that many consumers are poorly informed about, my investigation's objective is to assess whether the text descriptions that accompany wines convey information to consumers that affects their demand and, ultimately, influences the price charged. A journey through wines produced in the United States is undertaken to achieve this. First, I consider wines produced in New York. While the text correlates with price, once the varietal is controlled for, the statistical significance disappears. This suggests that the text is basically conveying the varietal being sold. To eliminate this effect, only one varietal is chosen in one state − Pinot Noir produced in Oregon. Again, within this sample, the text is correlated with price. Here, the statistical significance dampens when the numerical rating is controlled for. This suggests that some words and phrases convey high quality to consumers, while other descriptions do not. Thus, the textual description is not conveying anything more than the numerical score does. The final sample I consider zones in on not just only one varietal produced in one state, but also restricts to one particular numerical rating. Here I consider the most common rating of a wine produced in substantial volume in the United States: 90-rated, Cabernet Sauvignon wines produced in California. Focusing on this one subsample, I find that the text correlates with the prices charged. The effect cannot be explained away by the other information consumers may know at the time of sale. Thus, wine descriptions do indeed convey some information to consumers that affects their demand and, ultimately, price.

An important point made by Combris *et al.* (1997a) in their hedonic price analysis of wine is that while sensory characteristics affect the utility received from an experience good such as wine, which is reflected in its numerical rating, only observable characteristics are reflected in its price. This is due to the inability of a consumer of an experience good to correctly anticipate the sensory characteristics at the time of purchase. Text descriptions provide an avenue for consumers to collect some of this information. Therefore, it is not surprising that value this information. Furthermore, Ashenfelter (2008) illustrates that knowledge of the producer and the vintage explain a high percentage of the variation in the prices of wine. His data focuses, though, on the most elite, high-priced wines in the world. For my sample of wines, drunk more commonly by the typical consumer, the explanatory value of producer/region and vintage is more modest. Thus, there are other important variables that must be explaining the variation in prices observed. I show that the information provided in the wine's descriptions is closing this gap.

The primary contribution here is to identify the value of text to consumers in their purchasing of experience goods. Numerous future investigations that are not addressed here exist. For example, I use the text provided by professional wine reviewers. Producers can write their own descriptions. This can open up credence goods-related problems where the seller's diagnosis can be misrepresented (Bester and Dahm, 2018). It is unclear which is more informative to consumers. Relatedly, producers can choose which words to use in their descriptions. Presumably, the methods outlined here could guide such an effort. Furthermore, a retailer can select whether to post a description, and which to post. A prominent example of this is movie

reviews included in an advertisement or critic comments printed on a book jacket. The choice to not post can be informative (Bederson *et al.*, 2018), and this tradeoff is not considered here. I view the work presented as an important first step into using topic modeling approaches developed in computer science to extend our understanding of how markets function.

9 References

Ali, Hela Hadju, Sebastien Lecocq, and Michael Visser (2008), The Impact of Gurus: Parker Grades and *En Primeur* Wine Prices, *Economic Journal* 118(529): F158-F173.

Anderson, Michael and Jeremy Magruder (2012), Learning from the Crowd: Regression Discontinutity Estimates of the Effects of an Online Review Database, *Economic Journal* 122(563): 957-989.

Ashenfelter, Orley (2008), Predicting the Quality and Prices of Bordeaux Wine, *Economic Journal* 118(529): F174-F184.

Ashenfelter, Orley and Gregory V. Jones (2013), The Demand for Expert Opinion: Bordeaux Wine, *Journal of Wine Economics* 8(3): 285-293.

Bederson, Benjamin B., Ginger Zhe Jin, Phillip Leslie, Alexander J. Quinn, and Ben Zou (2018), Incomplete Disclosure: Evidence of Signaling and Countersignaling, *American Economic Journal: Microeconomics* 10(1): 41-66.

Bester, Helmut and Matthias Dahm (2018), Credence Goods, Costly Diagnosis, adn Subjective Evaluation, *Economic Journal* 128(611): 1367-1394.

Blei, David M., Andrew Y. Ng, and Michael L. Jordan (2003), Latent Dirichlet Allocation, *Journal of Machine Learning* 3: 993-1022.

Brown, Alexander L., Colin F. Camerer, and Dan Lovallo (2012), To Review or Not to Review? Limited Strategic Thinking at the Movie Box Office, *American Economic Journal: Microeconomics* 4(2): 1-26.

Combris, Pierre, Sebastien Lacocq, and Michael Visser (1997a), Estimation of a Hedonic Price Equation for Bordeaux Wine: Does Quality Matter?, *Economic Journal* 107:390-402.

Combris, Pierre, Sebastien Lacocq, and Michael Visser (1997b), Estimation of a Hedonic Price Equation for Burgundy Wine, *Applied Economics* 32: 961-967.

Crozet, Matthieu, Keith Hood, and Thierry Mayer (2012), Quality Sorting and Trade: Firm-Level Evidence for French Wine, *Review of Economic Studies* 79(2): 609-617.

Dubois, Pierre and Celine Nauges (2010), Identifying the Effect of Unobserved Quality and Expert Reviews in the Pricing of Experience Goods: Empirical Application on Bordeaux Wine, *International Journal of Industrial Organization* 28(3): 205-212.

Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence (2017), The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation, *Journal of Accounting and Economics* 64(2-3): 221-245.

Gentzkow, Matthew and Jesse M. Shapiro (2010), What Drives Media Slant? Evidence from U.S. Daily Newspapers, *Econometrica* 78(1): 35-71.

Gergaud, Olivier and Victor Ginsburgh (2008), Natural Endowments, Production Technologies and the Quality of Wines in Bordeaux: Does Terroir Matter?, *Economic Journal* 118(529): F142-F157.

Ginsburgh, Victor A. and Jan C. van Ours (2003), Expert Opinion and Compensation: Evidence from a Musical Competition, *American Economic Review* 93(1): 289-296.

Goldstein, Robin, Johan Almenberg, Anna Dreber, John W. Emerson, Alexis Herschkowitsch, and Jacob Katz (2008), Do More Expensive Wines Taste Better? Evidence from a Large Sample of Blind Tastings, *Journal of Wine Economics* 3(1): 1-9.

Grimmer, Justin (2010), A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases, *Political Analysis* 18(1): 1-35.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2018), Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach, *Quarterly Journal of Economics* 133(2): 801-870.

Hilger, James, Greg Rafert, and Sofia Villas-Boas (2011), Expert Opinion and the Demand for Experience Goods: An Experimental Approach in the Retail Wine Market, *Review of Economics and Statistics* 93(4): 1289-1296.

Jin, Ginger Zhe and Phillip Leslie (2003), The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards, *Quarterly Journal of Economics* 118(2): 409-451.

King Gary, Patrick Lam, and Margaret E. Roberts (2017), Computer-Assisted Keyword and Document Set Discovery from Unstructured Text, *American Journal of Political Science* 61(4): 971-988.

Lecocq, Sebastien and Michael Visser (2006), What Determines Wine Prices: Objective vs. Sensory Characteristics, *Journal of Wine Economics* 1(1): 42-56.

Landon, Stuart and Constance E. Smith (1998a), Quality Expectations, Reputation and Price, *Southern Economic Journal* 64(3): 628-647.

Landon, Stuart and Constance E. Smith (1998b), The Use of Quality and Reputation Indicators by Consumers: The Case of Bordeaux Wine, *Journal of Consumer Policy* 20(3): 289-323.

McCannon, Bryan C. (2012), The Value of Multiple Reviews, *Applied Economics* 44(12): 1521-1525.

Moretti, Enrico (2011), Social Learning and Peer Effects in Consumption: Evidence from Movie Sales, *Review of Economic Studies* 78(1): 356-393.

Oczkowski, Edward (1994), A Hedonic Price Function for Australian Premium Wine, *Australian Journal of Agricultural Economics* 38(1): 93-110.

Oczkowski, Edward (2001), Hedonic Wine Price Functions and Measurement Error, *Economic Record* 77(239): 374-382.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev (2010), How to Analyze Political Attention with Minimal Assumptions and Costs, *American Journal of Political Science* 54(1): 209-228.

Ramirez, Carlos D. (2008), Wine Quality, Wine Prices, and the Weather: Is Napa "Different"?, *Journal of Wine Economics* 3(2): 114-131.

Ramirez, Carlos D. (2010), Do Tasting Notes Add Value? Evidence from Napa Wines, *Journal of Wine Economics* 5(1): 143-163.

Reinstein, David A. and Christopher M. Snyder (2005), The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics, *Journal of Industrial Economics* 53(1): 27-51.

Schamel, Gunter (2009), Dynamic Analysis of Brand and Regional Reputation: The Case of Wine, *Journal of Wine Economics* 4(1): 62-80.

Schwarz, Carlo (2018), ldagibbs: A Command for Topic Modeling in Stata using Latent Dirichlet Allocation, *Stata Journal*, forthcoming.

Tirunillai, Seshadri and Gerard J. Tellis (2014), Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation, *Journal of Marketing Research* 51(4): 463-479.

Weil, Roman L. (2007), Debunking Critics' Wine Words: Can Amateurs Distinguish the Smell of Asphalt from the Taste of Cherries?, *Journal of Wine Economics* 2(2): 136-144.

Yang, Nan, Jill J. McCluskey, and Carolyn Ross (2009), Willingness to Pay for Sensory Properties in Washington State Red Wines, *Journal of Wine Economics* 4(1): 81-93.